

データ解析の初期段階における TextilePlotの活用

(株) データサイエンスコンソーシアム

仲真弓

慶應義塾大学, 早稲田大学, (株) データサイエンスコンソーシアム

柴田里程

Introduction

TextilePlot (Kumasaka & Shibata, 2008)

…「水平性を基準にした平行（並行）座標プロット」
parallel coordinate plot

…各変数軸を平行に並べたプロット

データ：各国の主要農水産物の自給率

(出典：総務省統計局「世界の統計 2016」)

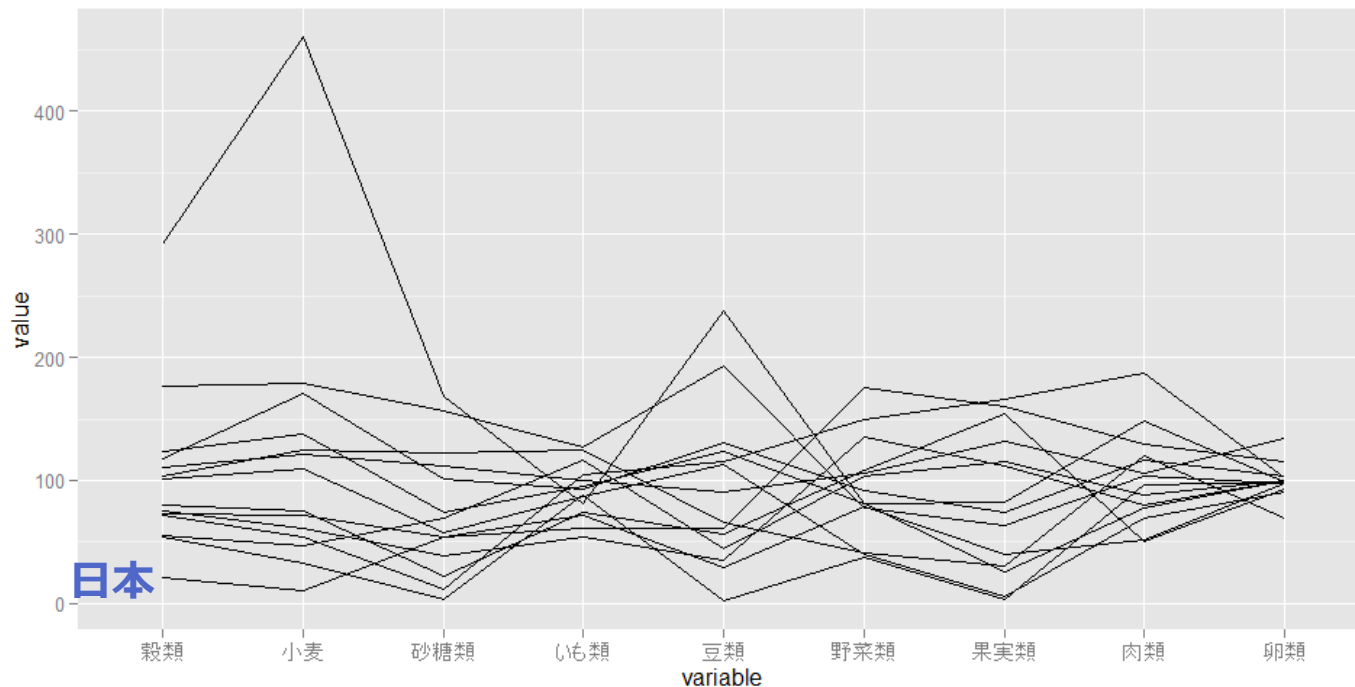
国(地域)	穀類	小麦	砂糖類	いも類	豆類	野菜類	果実類	肉類	卵類	魚介類
日本	20.9	10.1	53.8	71.2	29.2	79.3	40.1	51.3	98.1	57
トルコ	110.8	121.7	111.7	100.2	90.3	106	132.3	105.8	133.9	70.3
アメリカ合衆国	118.1	170.7	101	93	130.3	91.3	74.6	115.9	103.3	68.9
イギリス	101.7	109.7	57.5	86.9	113.4	40.1	5.3	69.9	90.7	50.1
イタリア	75.1	60.8	39.2	53.8	34.8	135.8	112.3	80.3	100.3	21.4
ギリシャ	79.5	74.8	21.7	74.6	56	107.9	153.7	51	93.5	34
スペイン	72.7	71.5	54.5	61.2	61.4	175.4	160.1	130	115.1	54.9
ドイツ	103.3	124.2	122.5	124.3	66.3	41.1	30.5	120.3	69.6	20.8
ノルウェー	53.9	32.9	2.9	88.4	1.8	37.3	3	96.9	98.3	198.6
フランス	176.6	178.8	156.1	126.7	192.5	77.7	63.6	103.1	97.5	29.8
ロシア	124	137.2	73.6	94.9	123.7	81.6	26.3	77.6	99.4	120.7
エジプト	54.7	46.5	69.4	116.1	44.2	104.1	115.5	87.9	99.7	76.1
オーストラリア	292.1	460.1	168.4	81.4	237.9	80.8	82.4	148.7	98.6	29.2
ニュージーランド	71.7	54.2	11.1	105.2	115.6	149.7	165.5	187.6	101.9	472.4

平行座標プロット

国(地域)	穀類	小麦	砂糖類	いも類	豆類	野菜類	果実類	肉類	卵類	魚介類
日本	20.9	10.1	53.8	71.2	29.2	79.3	40.1	51.3	98.1	57
トルコ	110.8	121.7	111.7	100.2	90.3	106	132.3	105.8	133.9	70.3
:	:	:	:	:	:	:	:	:	:	:



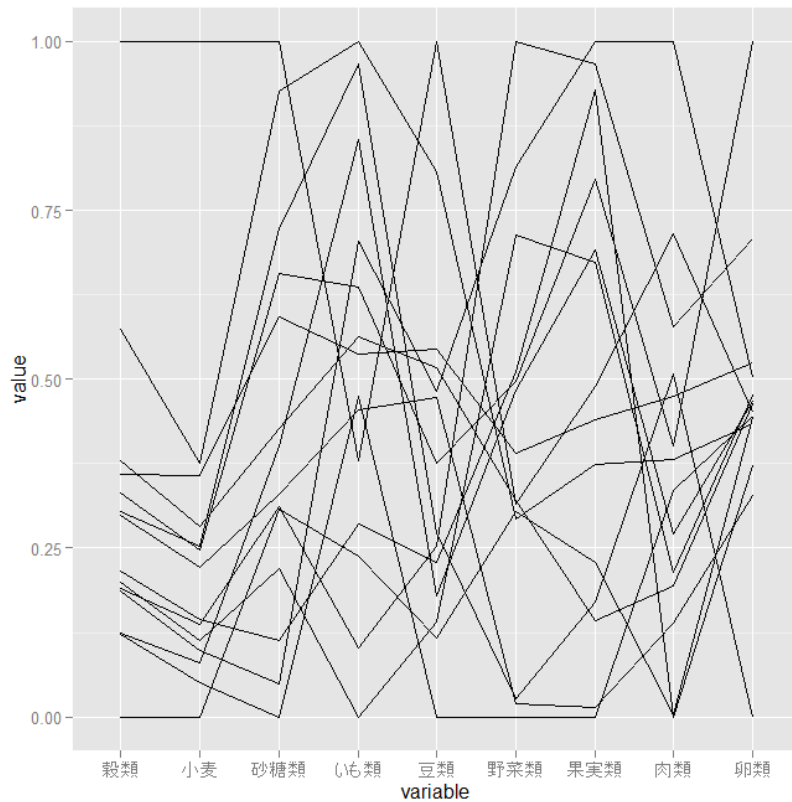
平行座標プロット (parallel coordinate plot)



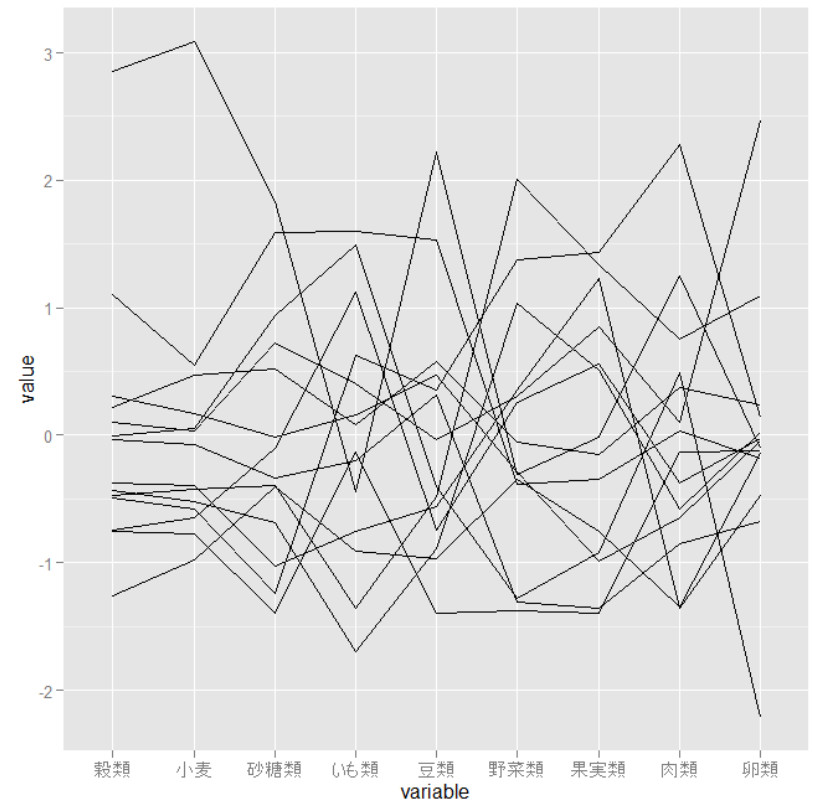
平行座標プロット

様々な基準による平行座標プロット

軸のスケールを各軸の最大・最小



軸ごとに平均・標準偏差で基準化



TextilePlot

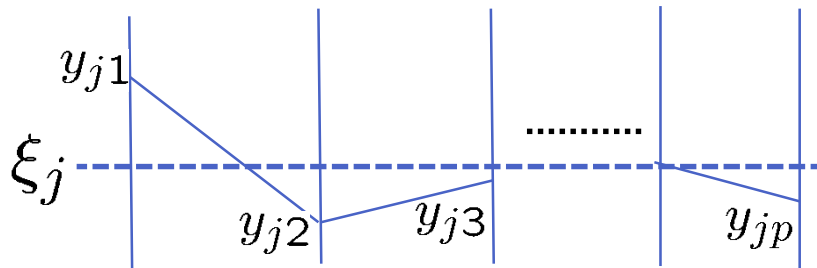
TextilePlot…「**水平性を基準**にした平行座標プロット」

$X = (x_1, x_2, \dots, x_p)$: データ行列

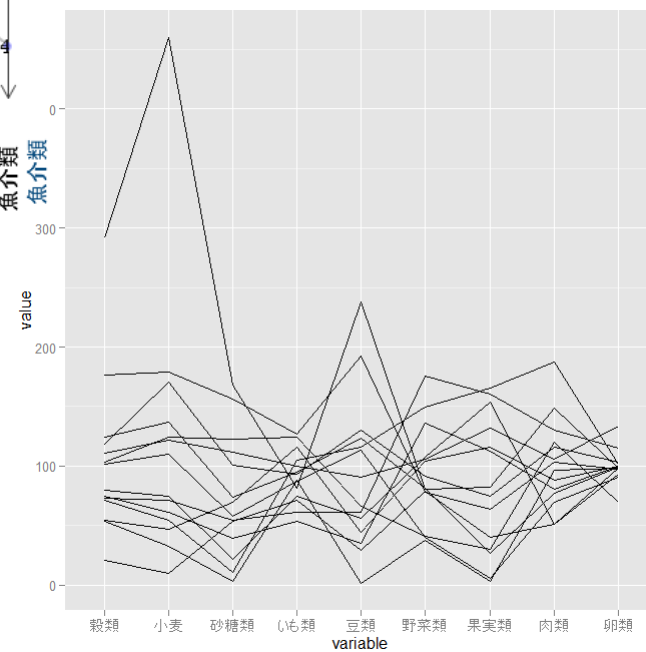
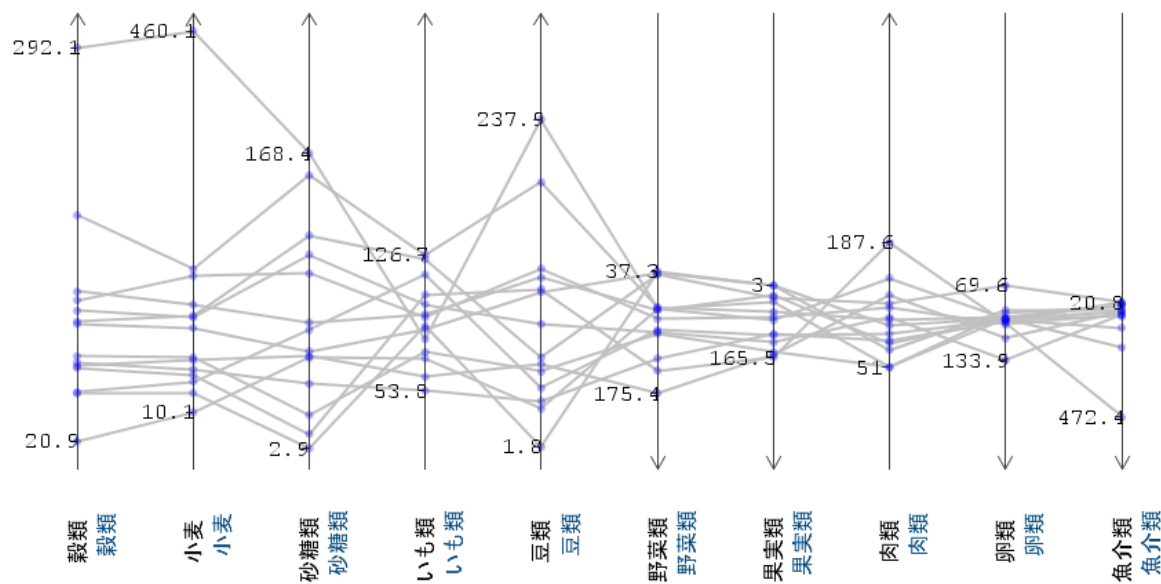
$$x_j = (x_{1j}, x_{2j}, \dots, x_{nj})^\top \quad \longrightarrow \quad \begin{aligned} y_j &= (y_{1j}, y_{2j}, \dots, y_{nj})^\top \\ &= \alpha_j \mathbf{1} + \beta_j x_j \end{aligned}$$

$$\sum_{j=1}^p \|y_j - \xi\|^2 \rightarrow \min$$

一つ一つの横線がなるべく水平になるように



TextilePlot



TextilePlot

- ノットがある…その変量が, 他の全てと直交 (無相関)
- 二つの軸について横線が平行…その二つの変量に線形関係がある (相関が1か-1)
- 因子変量, 順序因子変量でも, 同じ「水平性」を基準に座標を定められる

理論的な面での様々な可能性

- 全て数値変量の場合 …→ 主成分分析 (PCA) (Kumasaka et al., 2010)
- 全て因子変量の場合 …→ 多重コレスポネンス分析 (MCA) (Kumasaka et al., 2010)
- 座標の集合 $T_{n,p} = \{(y_1, y_2, \dots, y_p)\}$ への幾何学的アプローチ (Sei & Tanaka, 2015)
- OGI (客観的総合指数, Objective General Index)との関係 (Sei, 2016)

本報告では, 「データ解析の初期段階での活用」に注目

データ解析の初期段階での活用

- ① 簡単にデータの特徴を見ることができる
- ② 因子変量も同じ基準で、簡単にデータの特徴を見ることができる
- ③ 因子変量のレベルの分類のための予備解析ができる

データ解析の初期段階での活用

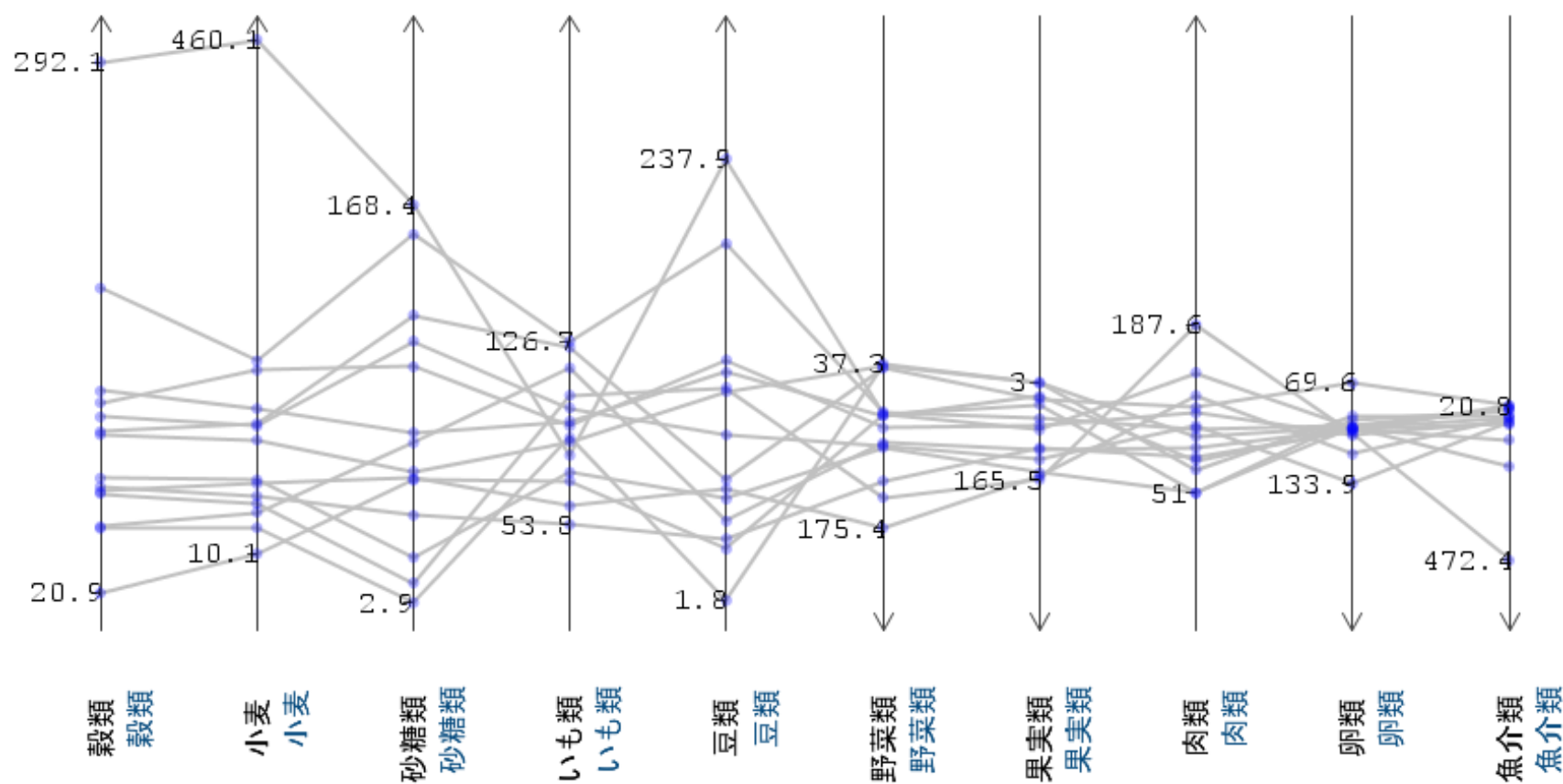
- ① 簡単にデータの特徴を見ることができる
→ 主要農水産物の自給率データ
- ② 因子変量も同じ基準で、簡単にデータの特徴を見ることができる
→ 気象データ
- ③ 因子変量のレベルの分類のための予備解析ができる
→ 不動産投資信託（REIT）データ

① 簡単にデータの特徴を見ることができる

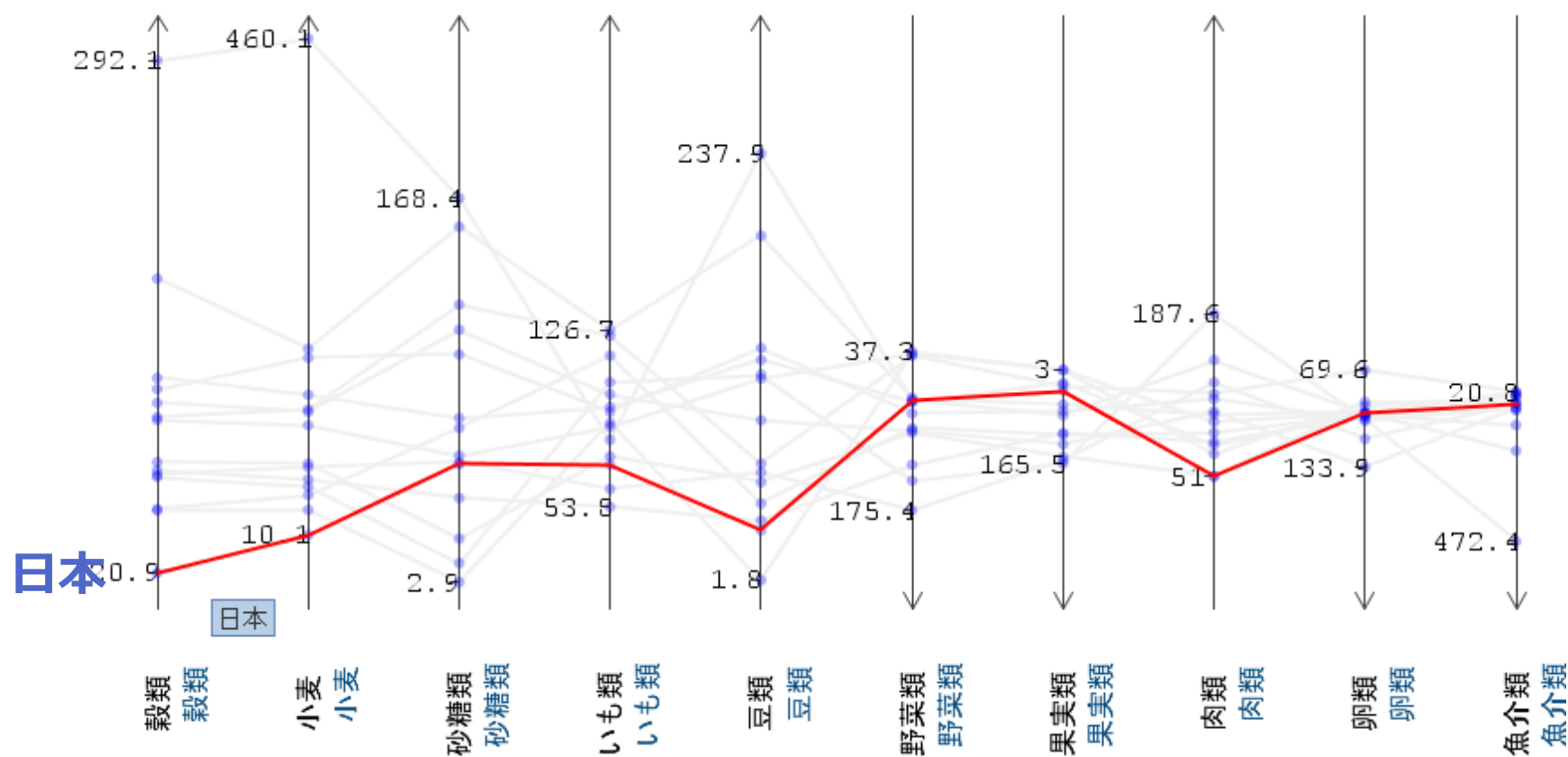
データ：主要農水産物の自給率（総務省統計局「世界の統計 2016」より）

国(地域)	穀類	小麦	砂糖類	いも類	豆類	野菜類	果実類	肉類	卵類	魚介類
日本	20.9	10.1	53.8	71.2	29.2	79.3	40.1	51.3	98.1	57
トルコ	110.8	121.7	111.7	100.2	90.3	106	132.3	105.8	133.9	70.3
アメリカ合衆国	118.1	170.7	101	93	130.3	91.3	74.6	115.9	103.3	68.9
イギリス	101.7	109.7	57.5	86.9	113.4	40.1	5.3	69.9	90.7	50.1
イタリア	75.1	60.8	39.2	53.8	34.8	135.8	112.3	80.3	100.3	21.4
ギリシャ	79.5	74.8	21.7	74.6	56	107.9	153.7	51	93.5	34
スペイン	72.7	71.5	54.5	61.2	61.4	175.4	160.1	130	115.1	54.9
ドイツ	103.3	124.2	122.5	124.3	66.3	41.1	30.5	120.3	69.6	20.8
ノルウェー	53.9	32.9	2.9	88.4	1.8	37.3	3	96.9	98.3	198.6
フランス	176.6	178.8	156.1	126.7	192.5	77.7	63.6	103.1	97.5	29.8
ロシア	124	137.2	73.6	94.9	123.7	81.6	26.3	77.6	99.4	120.7
エジプト	54.7	46.5	69.4	116.1	44.2	104.1	115.5	87.9	99.7	76.1
オーストラリア	292.1	460.1	168.4	81.4	237.9	80.8	82.4	148.7	98.6	29.2
ニュージーランド	71.7	54.2	11.1	105.2	115.6	149.7	165.5	187.6	101.9	472.4

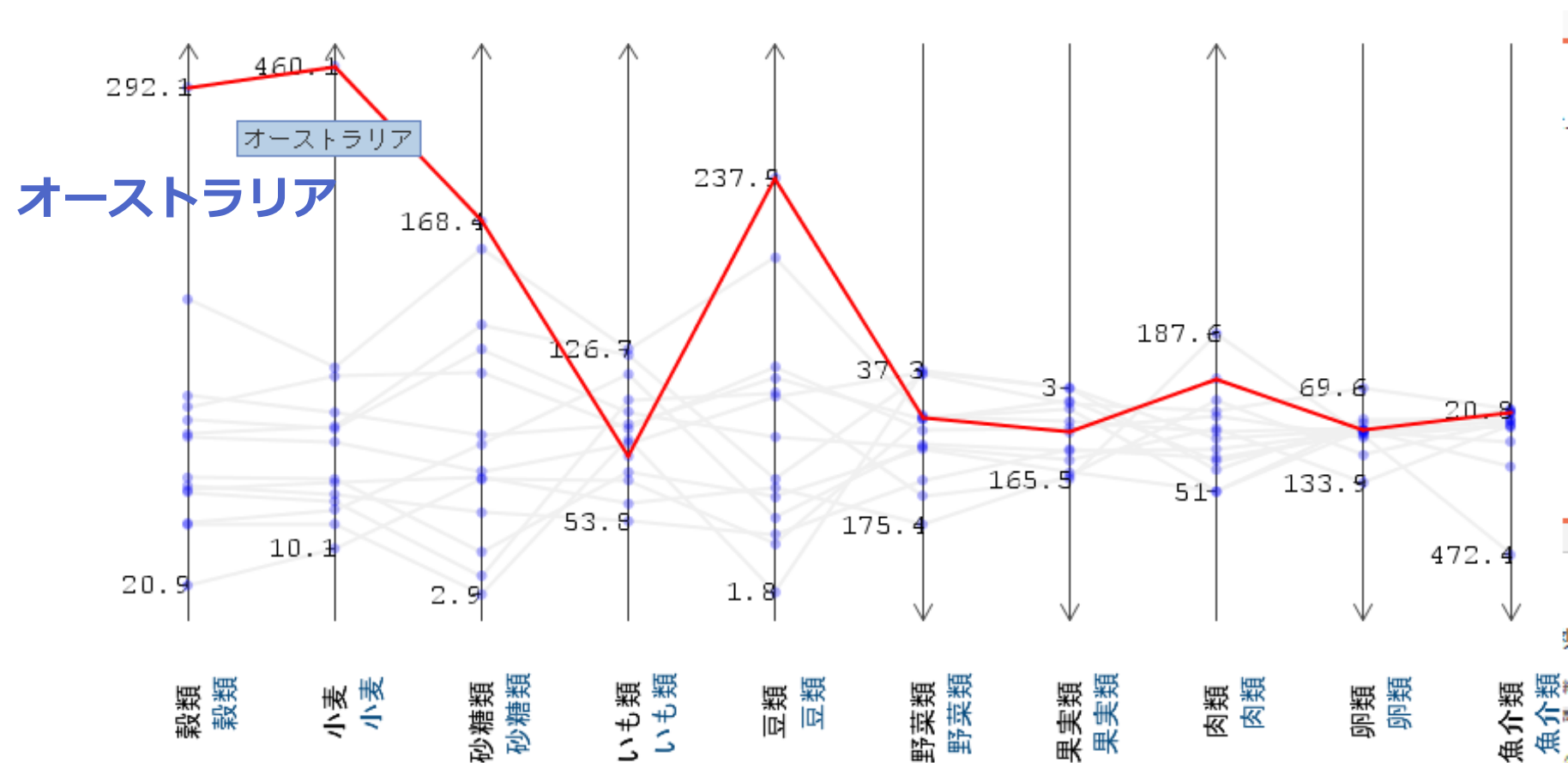
① 簡単にデータの特徴を見ることができる



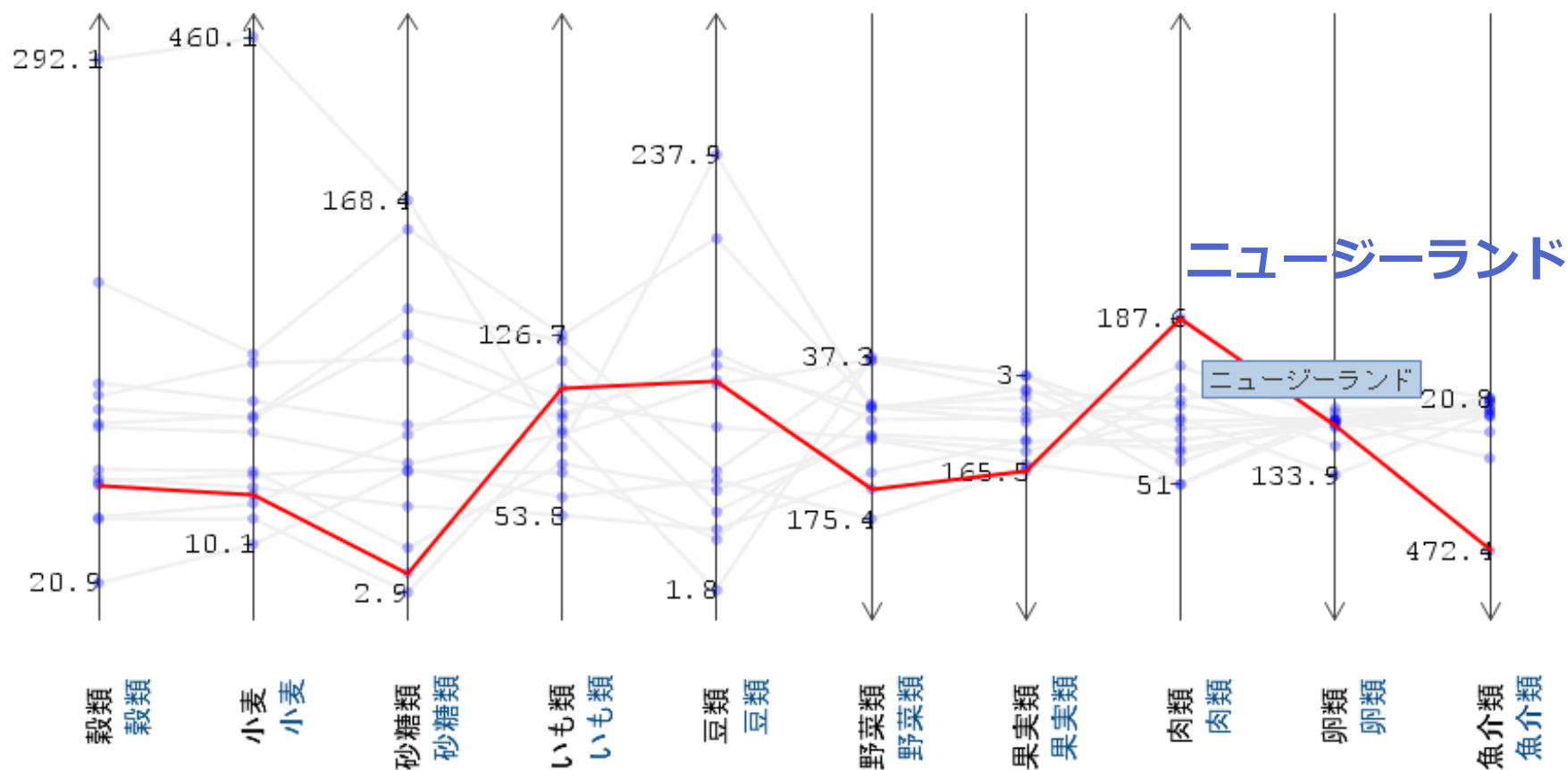
① 簡単にデータの特徴を見ることができる



① 簡単にデータの特徴を見ることができる



① 簡単にデータの特徴を見ることができる



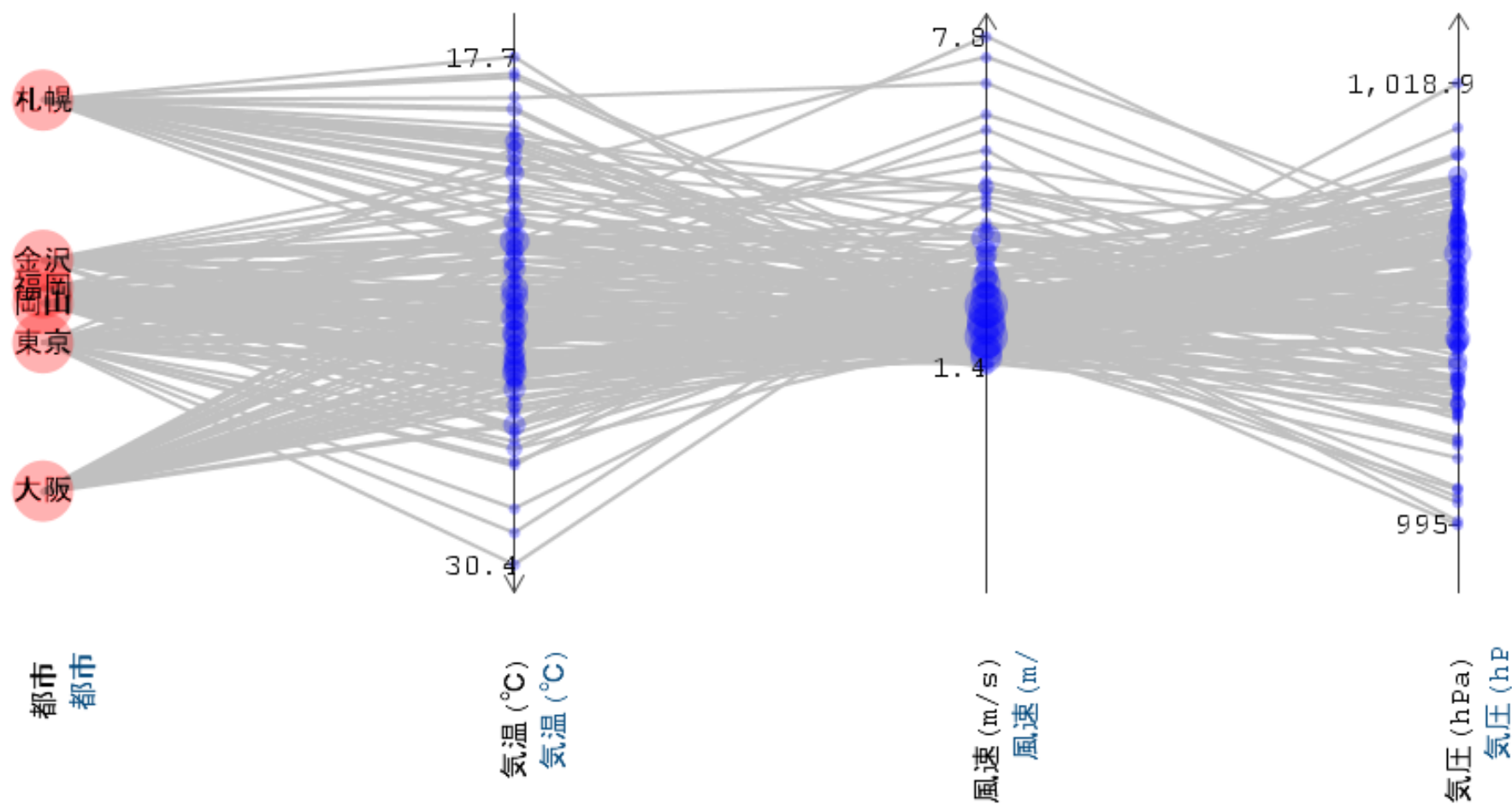
② 因子変量も同じ基準で見ることができる

データ：各都市の気象（気象庁のデータをもとに作成）

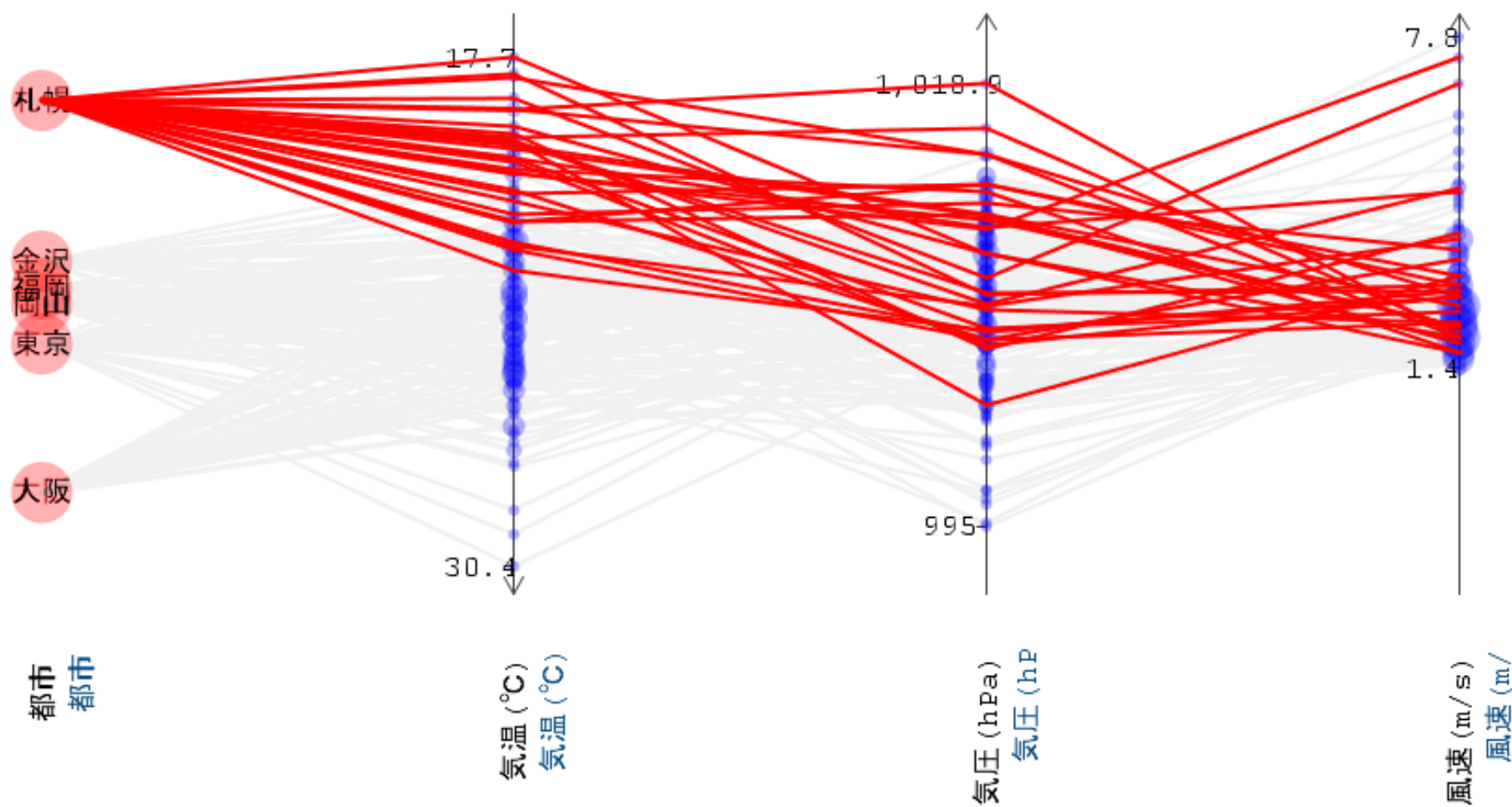
- ・ 2013-2015年の9月1日～10日の各日の平均の気温，風速，気圧のデータ
- ・ 対象の都市：札幌，東京，金沢，大阪，岡山，福岡

都市	気温(℃)	風速(m/s)	気圧(hPa)
金沢	23.9	3.6	1007
金沢	23.7	4.7	1008.6
金沢	24	2	1007.5
金沢	23.8	4.2	1006.9
：	：	：	：

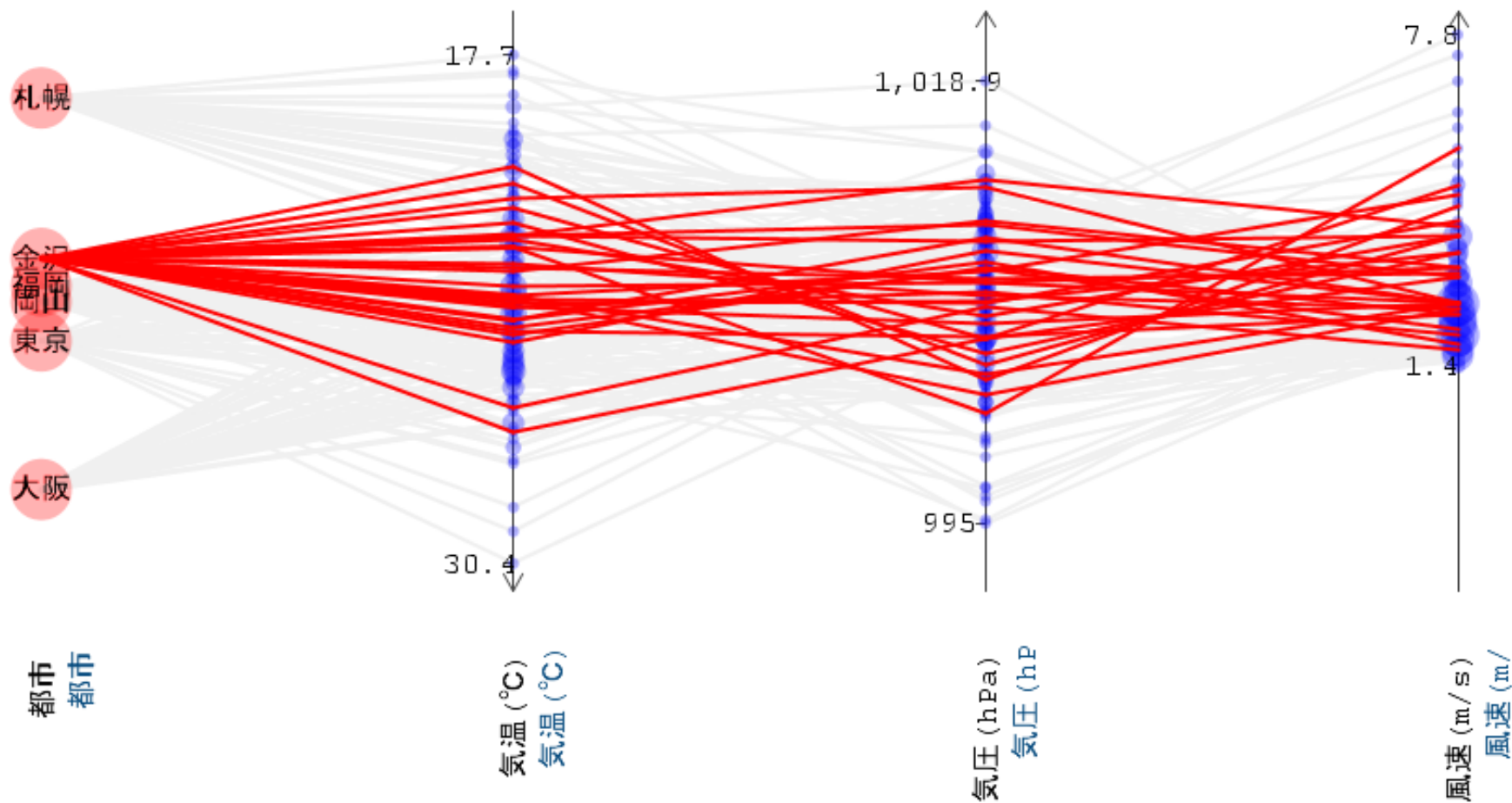
② 因子変量も同じ基準で見ることができる



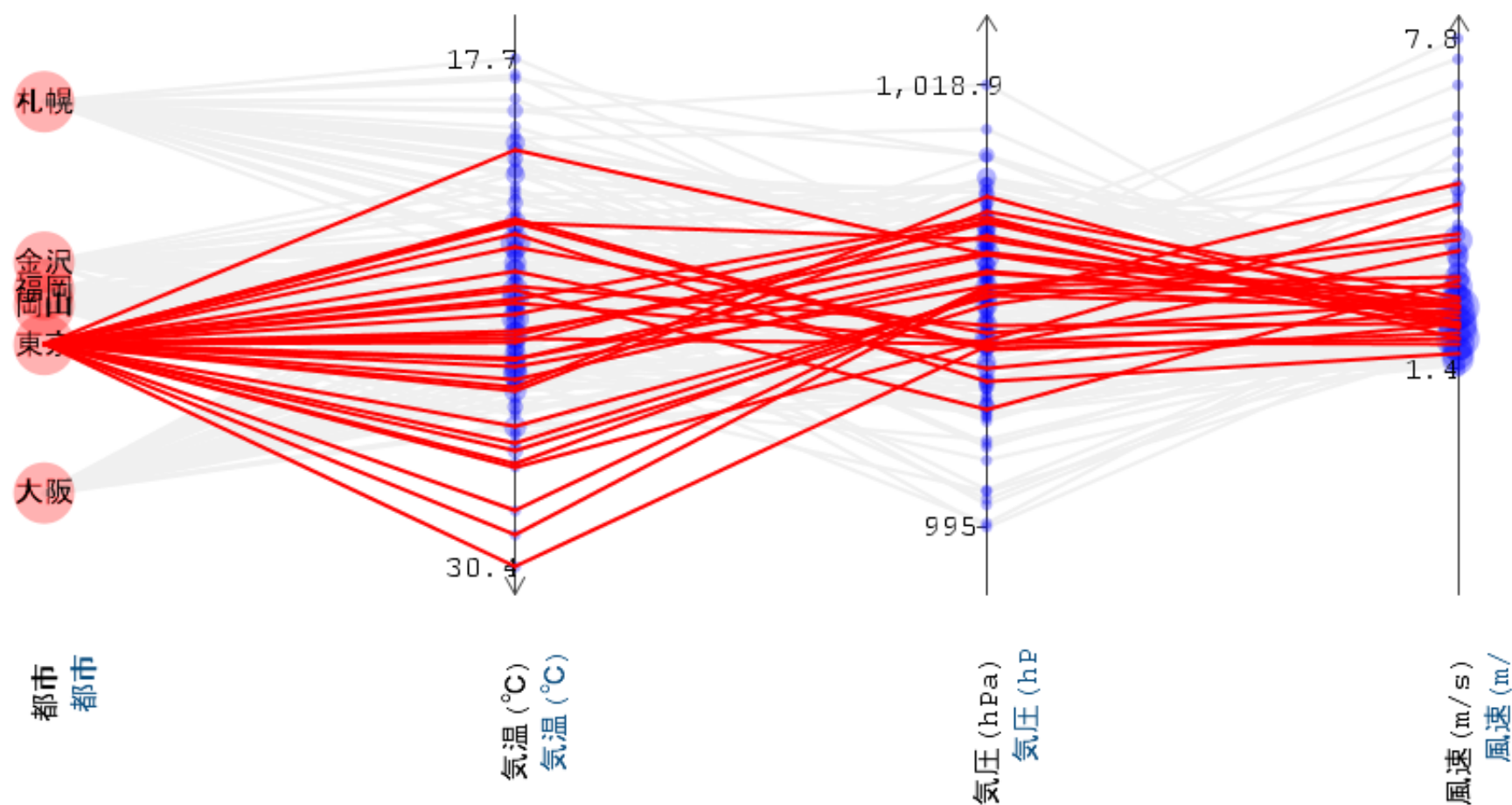
② 因子変量も同じ基準で見ることができる



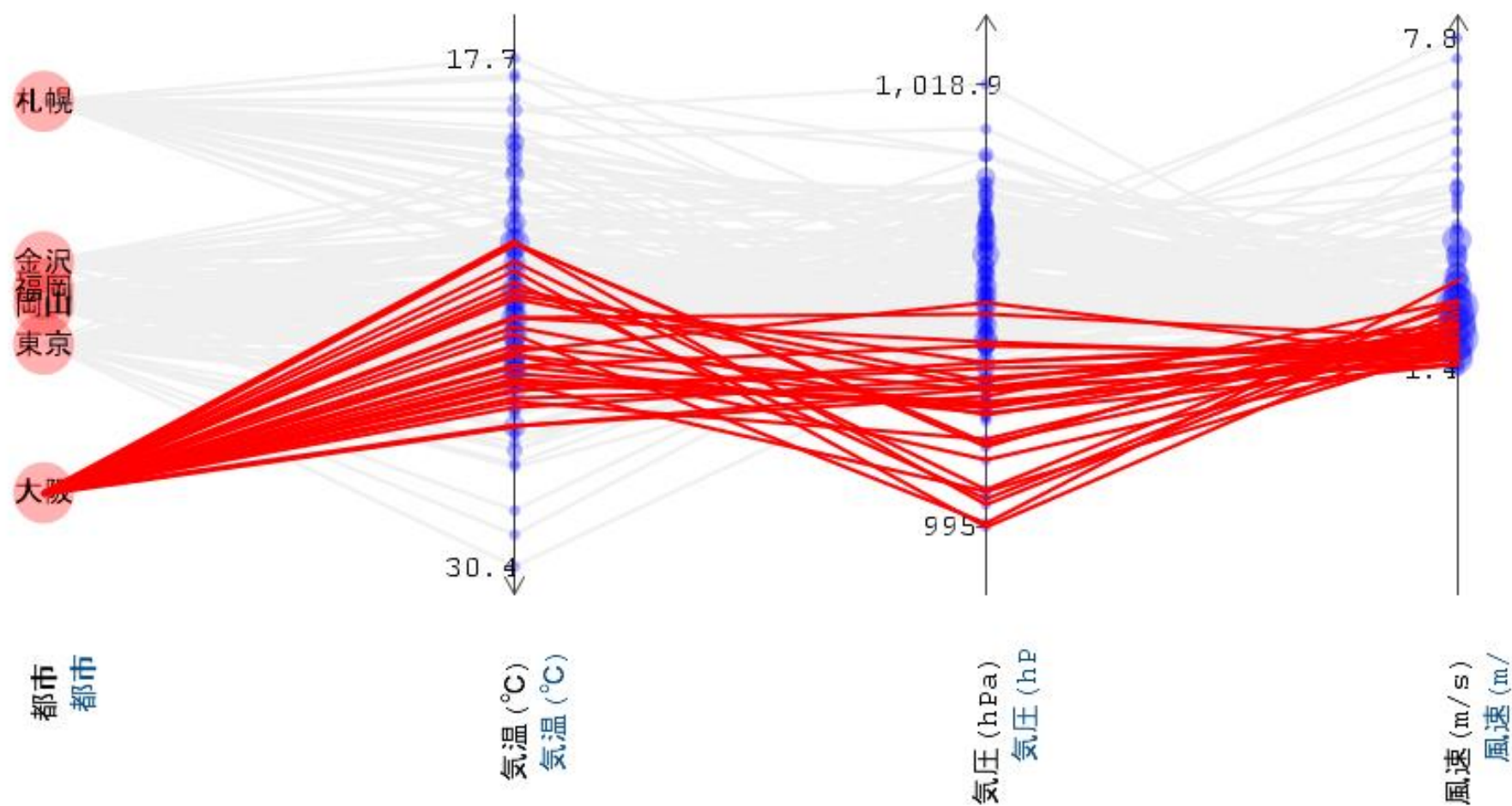
② 因子変量も同じ基準で見ることができる



② 因子変量も同じ基準で見ることができる



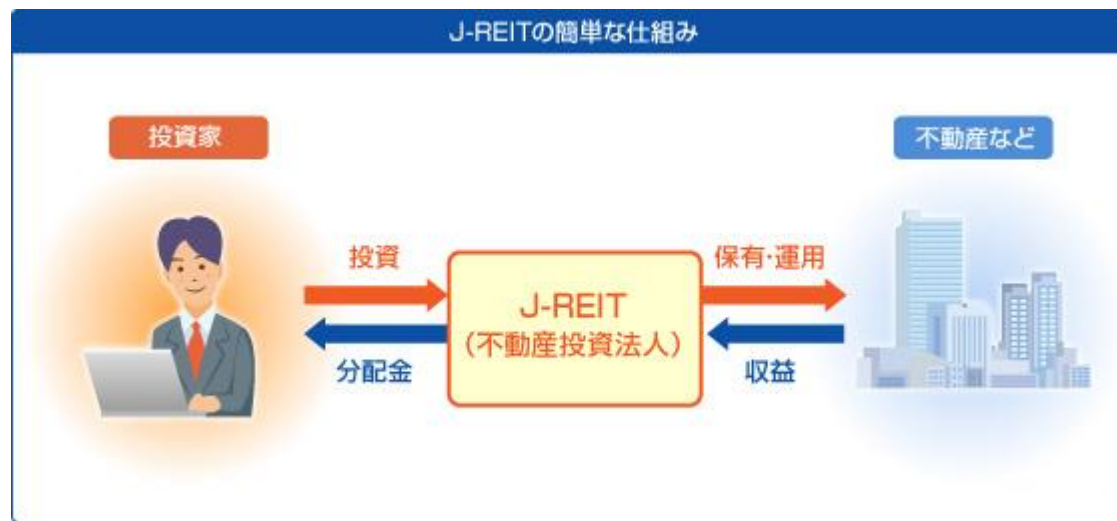
② 因子変量も同じ基準で見ることができる



③ 因子変量のレベルの分類のための予備解析ができる

データ：不動産投資信託（REIT）データ

- REIT (Real Estate Investment Trust) 不動産投資信託



一般社団法人 投資信託協会 <https://www.toushin.or.jp/reit/about/what/>

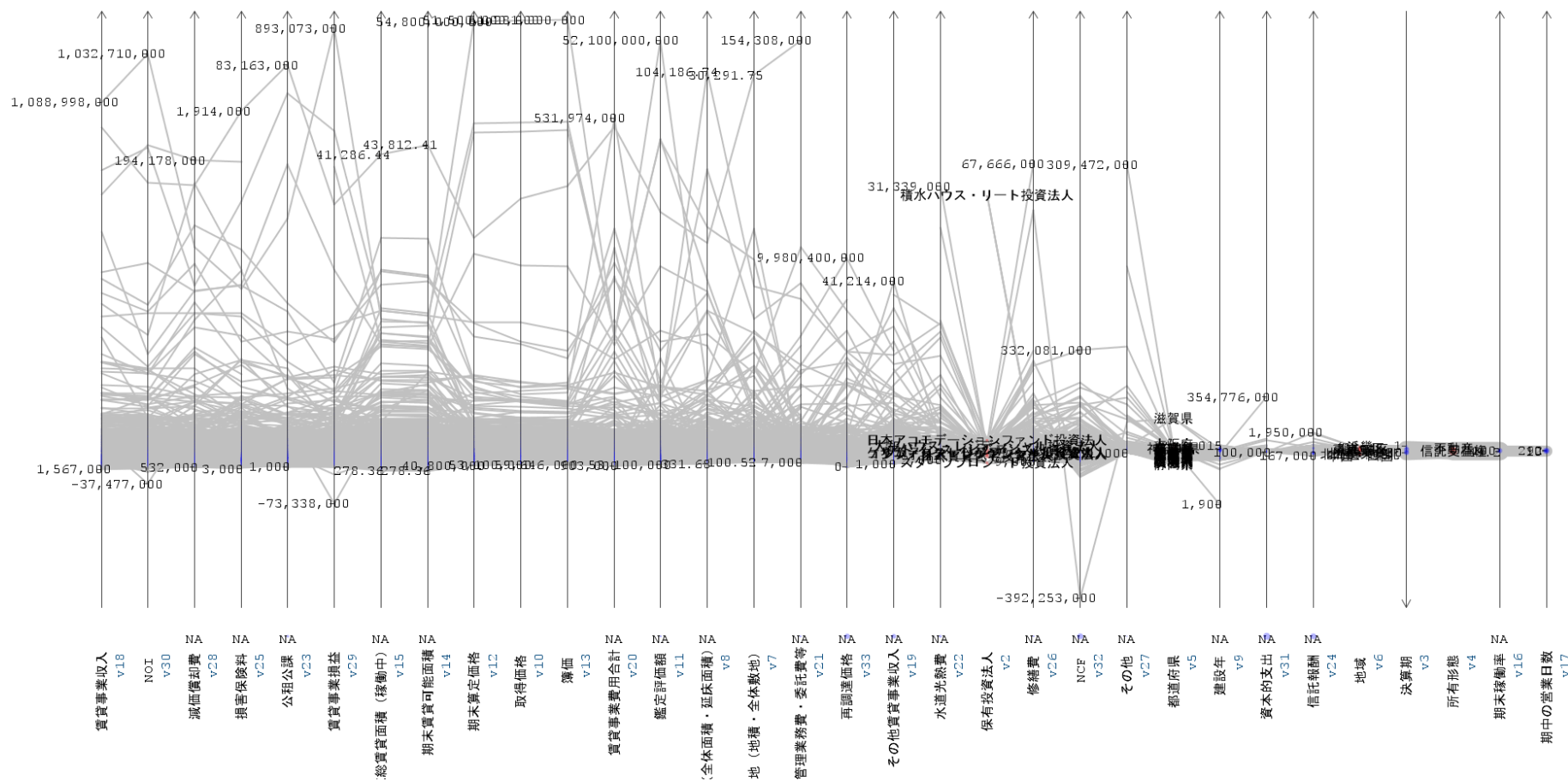
- J-REIT…56の不動産投資法人

③ 因子変量のレベルの分類のための予備解析ができる

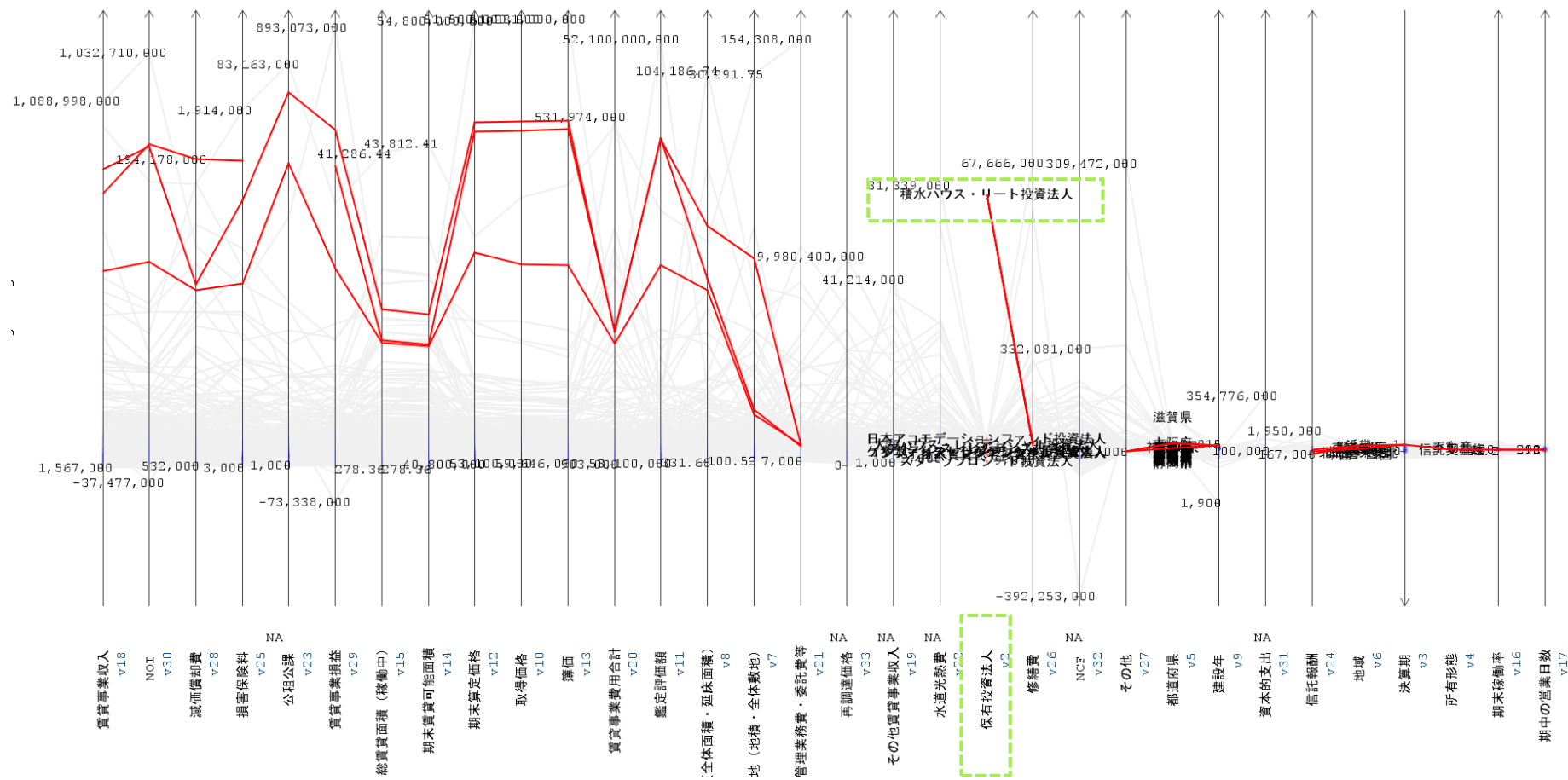
- データ：ARES J-REITデータベース
 - 対象：住宅主体系の9つの投資法人
 - 計1030個の直近の決算時の不動産データ
(1つの投資法人につき, 4~251個の不動産を運用)
 - 投資法人 / 決算期 / 所有形態 / ... 32の変数

9つの投資法人は、どれとどれが似ているか？

③ 因子変量のレベルの分類のための予備解析ができる

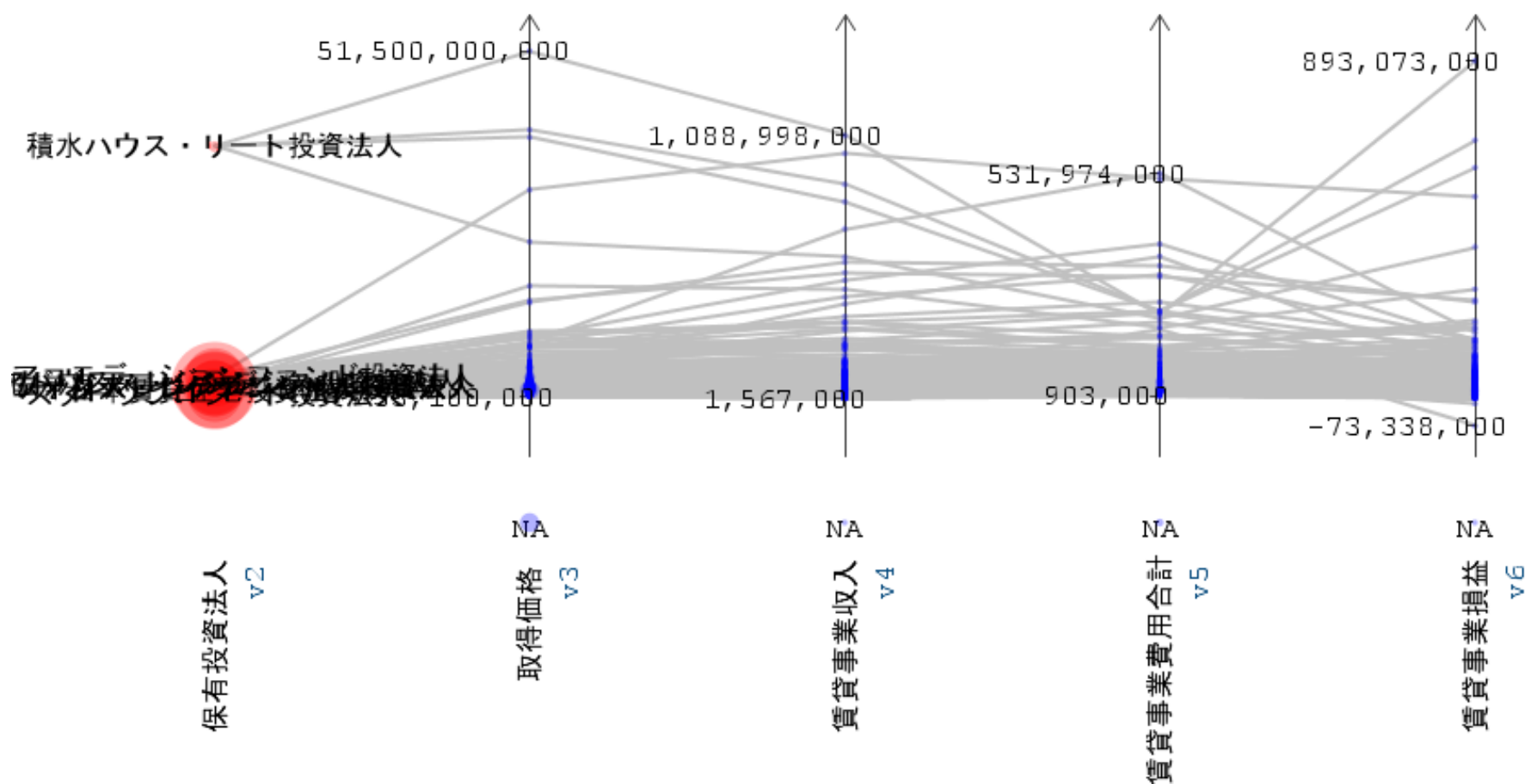


③ 因子変量のレベルの分類のための予備解析ができる



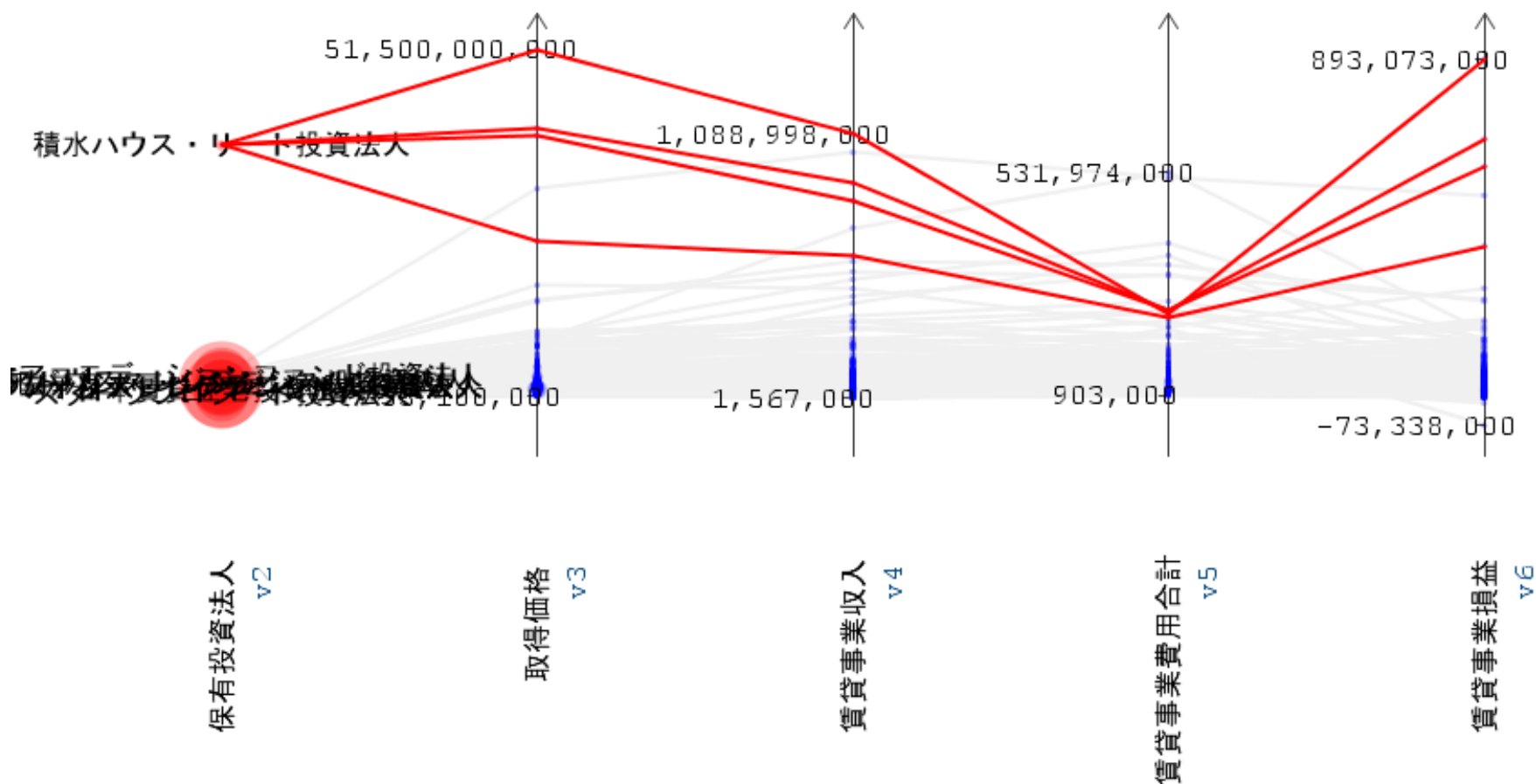
③ 因子変量のレベルの分類のための予備解析ができる

金額と関係のある変量に注目



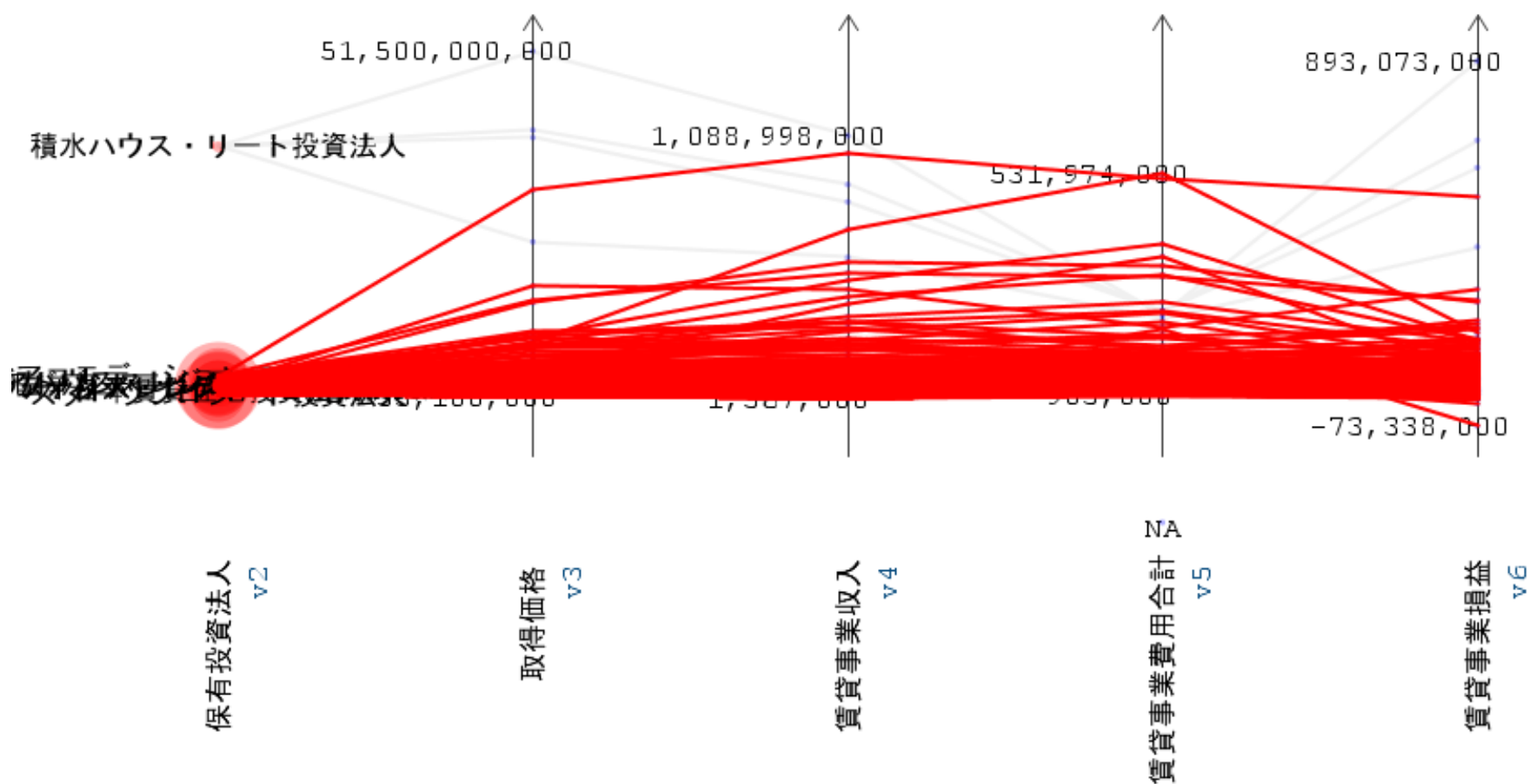
③ 因子変量のレベルの分類のための予備解析ができる

金額と関係のある変量に注目



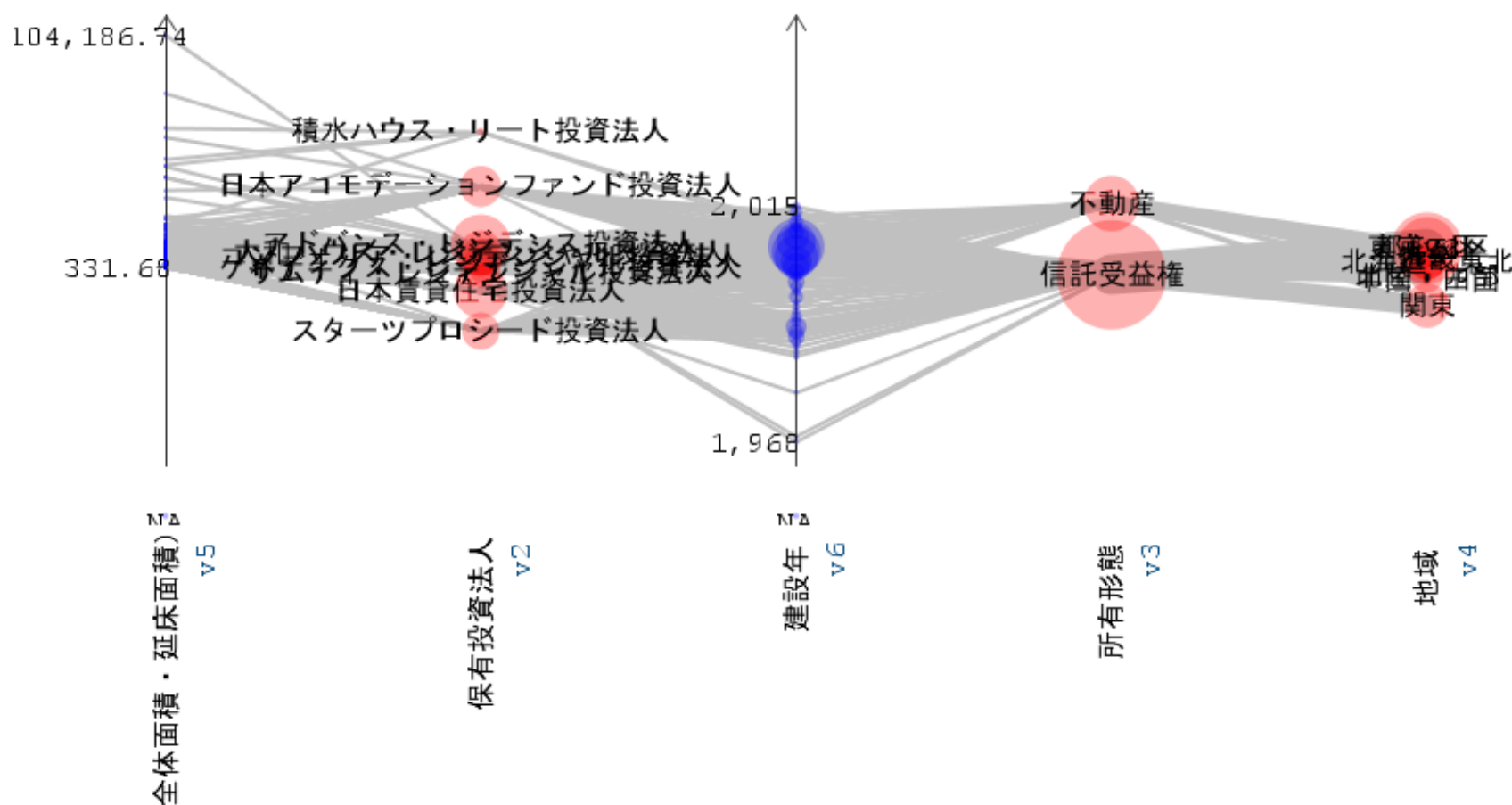
③ 因子変量のレベルの分類のための予備解析ができる

金額と関係のある変量に注目



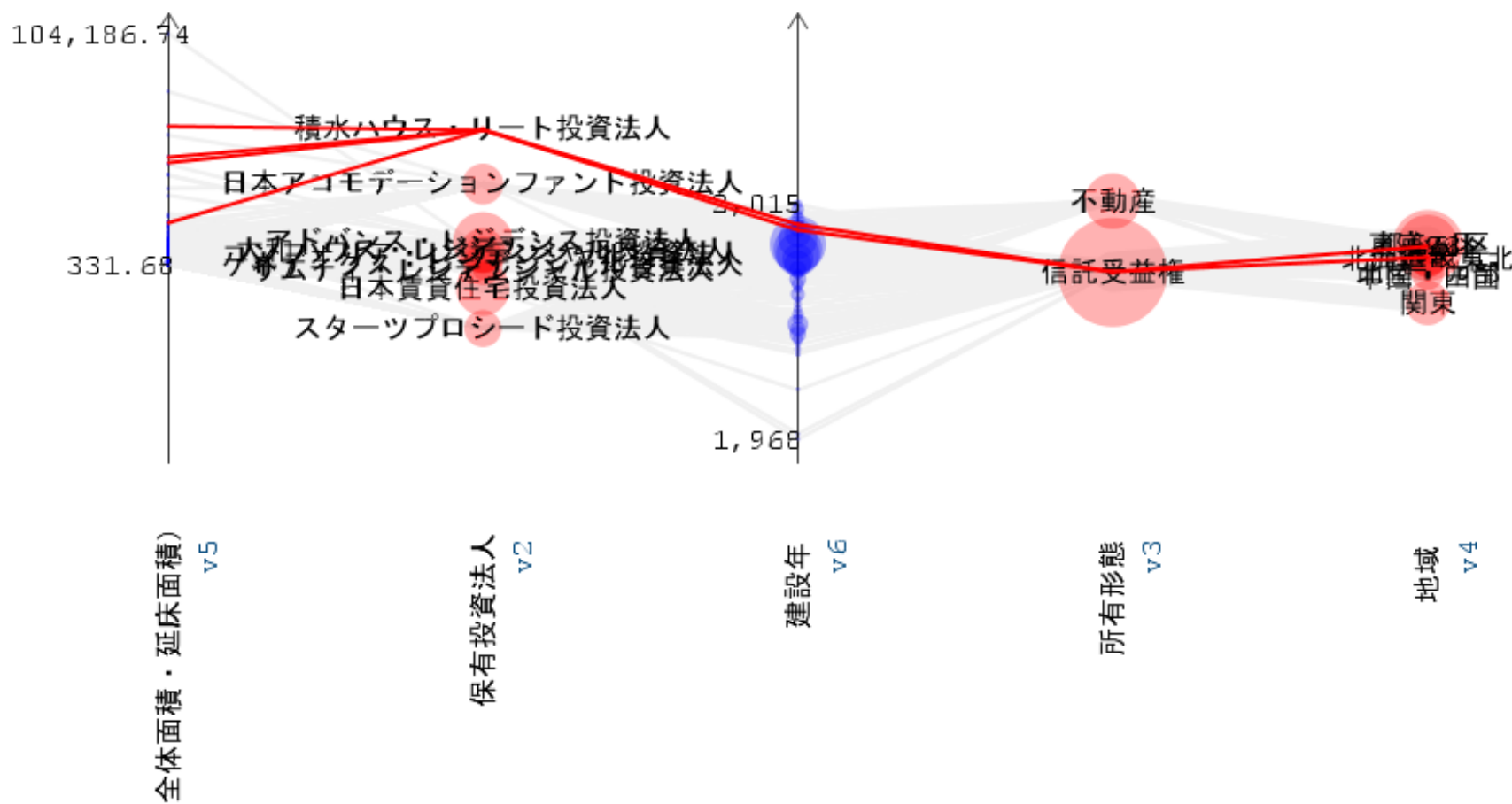
③ 因子変量のレベルの分類のための予備解析ができる

金額とは関係のない変量に注目



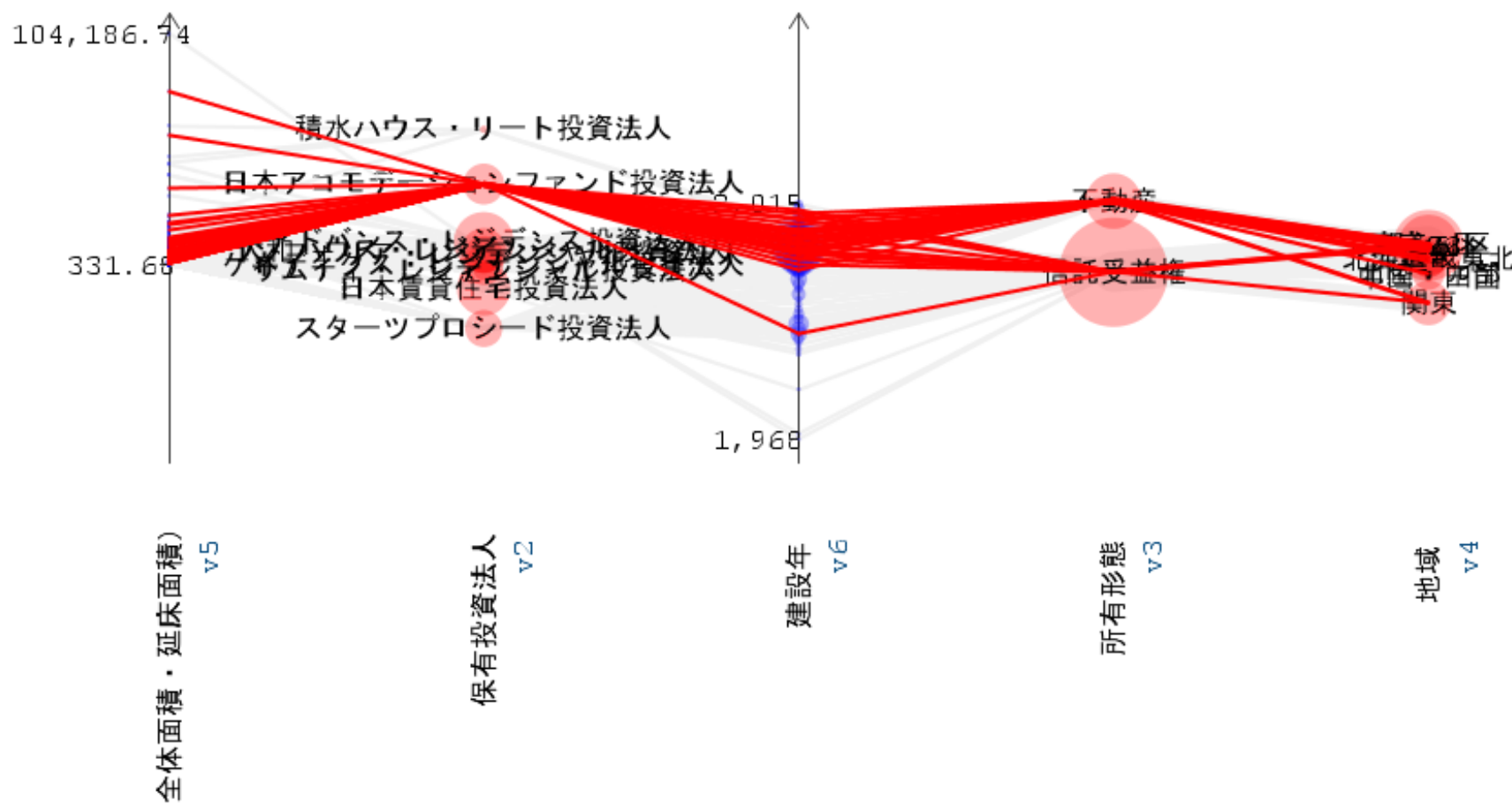
③ 因子変量のレベルの分類のための予備解析ができる

金額とは関係のない変量に注目



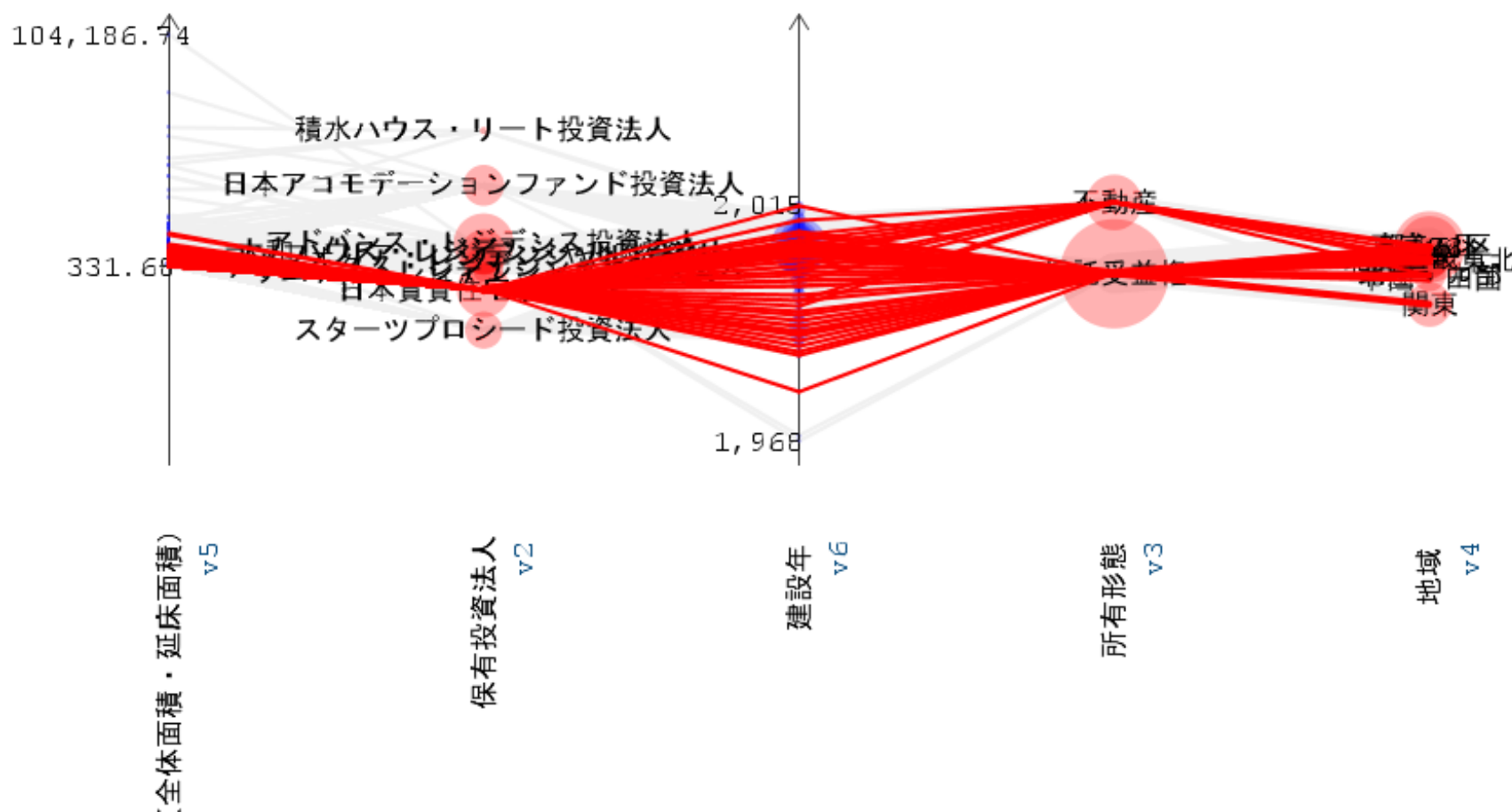
③ 因子変量のレベルの分類のための予備解析ができる

金額とは関係のない変量に注目



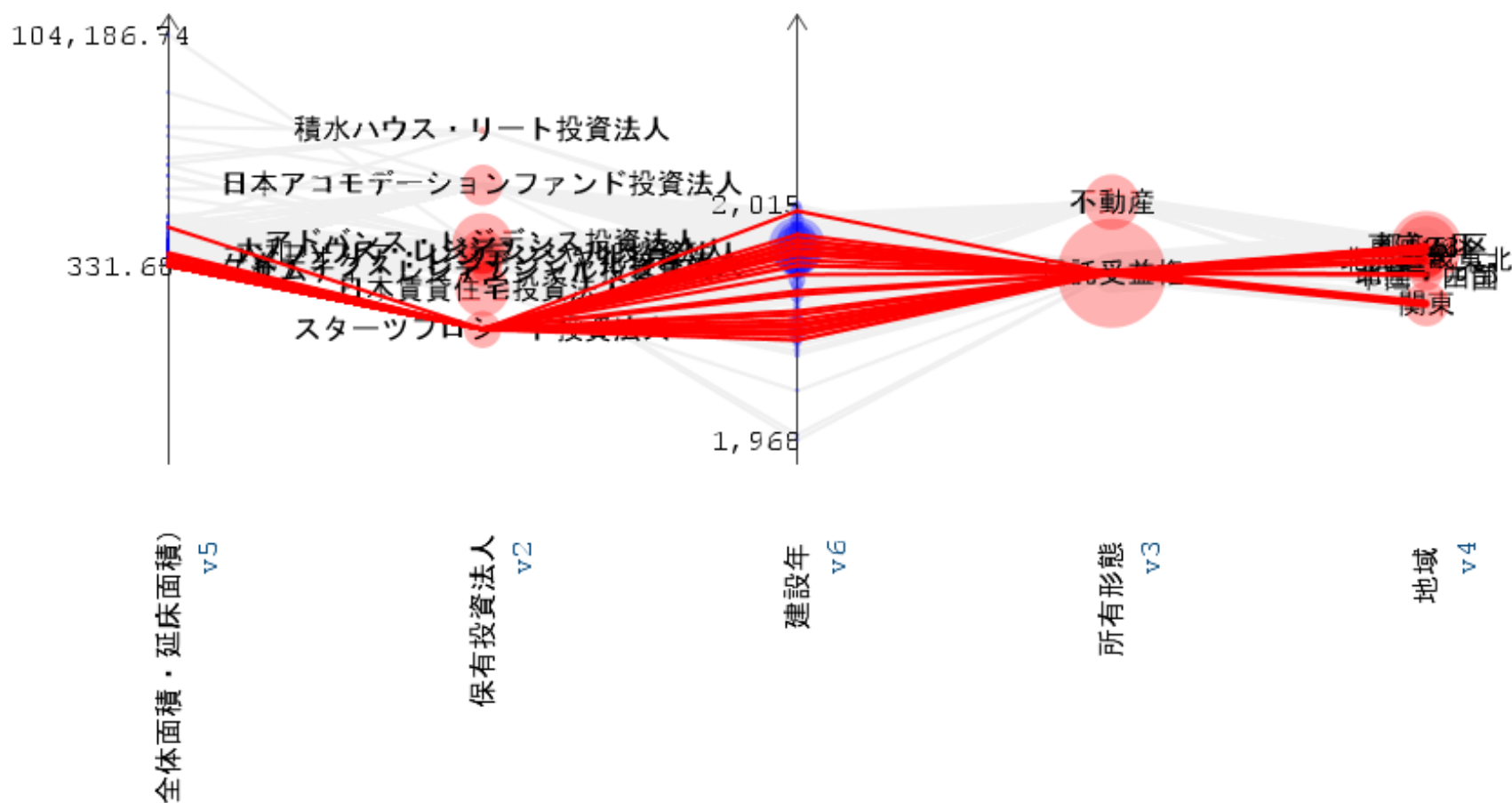
③ 因子変量のレベルの分類のための予備解析ができる

金額とは関係のない変量に注目



③ 因子変量のレベルの分類のための予備解析ができる

金額とは関係のない変量に注目



Summary

TextilePlot … 「水平性を基準にした平行（並行）座標プロット」

データ解析の初期段階での活用

- ① 簡単にデータの特徴を見ることができる…主要農水産物の自給率データ
- ② 因子変量も同じ基準で，簡単にデータの特徴を見ることができる…気象データ
- ③ 因子変量のレベルの分類のための予備解析ができる…不動産投資信託データ

+ 理論的にも様々な可能性

参考文献

Kumasaka, N. & Shibata, R. (2008). High-dimensional data visualisation: The textile plot. *Computational Statistics & Data Analysis*, 52(7), 3616-3644.

Kumasaka, N., Nakamura, Y., & Kamatani, N. (2010). The textile plot: a new linkage disequilibrium display of multiple-single nucleotide polymorphism genotype data. *PLoS One*, 5(4), e10207

Sei, T. & Tanaka, U. (2015). Geometric Properties of Textile Plot: Geometric Science of Information, *Lecture Notes in Computer Science*, 9389, 732-739.

Sei, T. (2016). An objective general index for multivariate ordered data. *Journal of Multivariate Analysis*, 147, 247-264.